

Proposal for Creation of Algorithm for Identification of Native Language of Author of Any Algorithmically-Translated Text Passage for Purpose of More Readily Identifying Propaganda Source

3 February 2022

Simon Edwards

Research Acceleration Initiative

Introduction

Both human and algorithm-translated text feature telltale clues as to the native language of the original author. Most but not all state propaganda is prepared by humans, but this does not preclude an analyst from ascertaining the language of origin of the author. While that much information can usually be inferred from the content of the propaganda, propaganda is sometimes crafted as part of an attempt to prompt the misattribution of the source, making it useful to be able to use language meta-analysis to determine the true language of origin of the author.

Abstract

Such an algorithm could be programmed to look for linguistic collocations specific to the language in question, noun-verb sequence, gendered language, excessively short sentences, run-on sentences, adjective overuse (Spanish and Russian,) failure to use conjunctions (Mandarin,) and the specific sorts of malapropisms used (as well as those malapropisms purposefully avoided by an author, just to name a few potential targets of the algorithm.

Conclusion

Identifying automatically translated text would be a trifling matter, however, expertly translated text is more challenging to analyze. The ability to use algorithms to identify expertly translated texts may prove important as it can allow for rapid detection of deployed propaganda and for counter-propaganda narratives to be constructed on a reactionary basis with improved response times. Critical is the accurate captioning of videos as the captioning process oftentimes introduces additional errors which would render any such algorithm moot.